Can Chinese Web Pages be Classified with English Data Source?

Xiao Ling[†] Gui-Rong Xue[†*] Wenyuan Dai[†] Yun Jiang[†] Qiang Yang[‡] Yong Yu[†]

[†]Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
{shawnling, grxue, dwyak, yunjiang, yyu}@apex.sjtu.edu.cn

[‡]Hong Kong University of Science and Technology, Clearway Bay, Kowloon, Hong Kong
qyang@cs.ust.hk

ABSTRACT

As the World Wide Web in China grows rapidly, mining knowledge in Chinese Web pages becomes more and more important. Mining Web information usually relies on the machine learning techniques which require a large amount of labeled data to train credible models. Although the number of Chinese Web pages increases quite fast, it still lacks Chinese labeled data. However, there are relatively sufficient English labeled Web pages. These labeled data, though in different linguistic representations, share a substantial amount of semantic information with Chinese ones, and can be utilized to help classify Chinese Web pages. In this paper, we propose an *information bottleneck* based approach to address this cross-language classification problem. Our algorithm first translates all the Chinese Web pages to English. Then, all the Web pages, including Chinese and English ones, are encoded through an information bottleneck which can allow only limited information to pass. Therefore, in order to retain as much useful information as possible, the common part between Chinese and English Web pages is inclined to be encoded to the same code (i.e. class label), which makes the cross-language classification accurate. We evaluated our approach using the Web pages collected from Open Directory Project (ODP). The experimental results show that our method significantly improves several existing supervised and semi-supervised classifiers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Cross-Language Classification, Information Bottleneck

1. INTRODUCTION

A dramatic development of Internet in China has been witnessed in the recent years. The number of Internet users in China now exceeds 160 millions and the Chinese Web pages are numbered in billions[1]. As the most commonly used language only second to English, Chinese is expected to enjoy such a rocketing increase in scale. Because the Web pages written in Chinese is becoming a major information source on the Internet, more research efforts are now devoted to organizing and mining the Chinese Web pages via Web mining techniques, such as Chinese blog mining [20] and query log analysis [19].

A potential problem in mining the Chinese Web pages is the lack of sufficient labeled data. As we know, classification requires a large amount of labeled training data. Generally speaking, the more labeled training data one can obtain, the better the classification accuracy and robustness are. Fortunately, due to many reasons, there exists a lot of labeled Web-page information in English, in particular in the machine learning community. Examples of these resources are Reuters-21578 [16], 20 Newsgroups [15], and Open Document Project [22]. It is thus useful and intriguing to fully utilize the labeled documents in English to help classify the Web pages in Chinese. This problem is called *cross-language Web-page classification*. In this paper, we address this important problem using a novel information theory based technique.

Although the training and test documents are in different languages, one can use a translation tool to help translate the test data sets in the English language, before a classifier trained on English pages can be applied. While such a method may be feasible, we observe that a simple-minded application of this method may result in serious problems because of the following reasons:

- First, due to the difference in language and culture, there exists a topic drift when we move from the English Web pages to the Chinese Web pages. This corresponds to the situation in machine learning where the training and test data have different distributions in terms of the class labels. This topic-imbalanced problem needs to be overcome in our research.

- Second, due to the errors introduced in the translation process, there may be different kinds of errors in the translated text. For example, some errors may result from Chinese phrase segmentation, others are due to ambiguities introduced by a dictionary. This translation noise problem must be addressed effectively.

*Gui-Rong Xue is the corresponding author.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

[1]According to the report by CNNIC in January 2007: http://www.cnnic.net.cn/uploadfiles/pdf/2007/2/13/95522.pdf

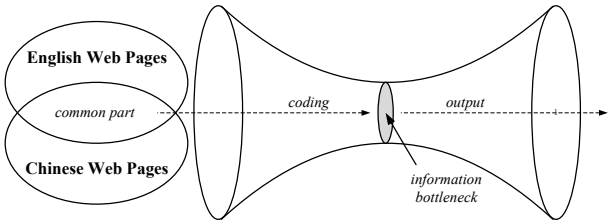

Figure 1: The model of our information bottleneck cross-language classifier.

- Finally, the feature spaces of the English and Chinese Web pages may be different, resulting in a situation where the training and test data may have different feature sets. We therefore must be innovative in our solution to this problem by carefully extracting the common-semantic parts of the two data sets, and use these parts as a bridge to propagate the class labels.

To solve the above problems, we develop a novel approach for classifying the Web pages in Chinese using the training documents in English. Our key observation is that despite the above listed problems, linguistically, the pages in Chinese and English may share the same semantic information, although they are in different representation forms, i.e., Chinese characters and English words, respectively. This is reasonable, because people with good command of both English and Chinese can convey the same information in both languages. Also, it is noted that the same meaning might be expressed in different ways due to the cultural and linguistic differences.

Based on the observations, we propose to tackle the cross-language classification problem using *information bottleneck* theory [29]. Figure 1 illustrates our idea intuitively. The training and translated target texts are encoded together, allowing all the information to be put through a "bottleneck" and represented by a limited number of codewords (i.e. labels in the classification problem). Information bottleneck based approach can maintain most of the common information and disregard the irrelevant information. Thus we can approximate the ideal situation where similar training and translated test pages, which is the common part, are encoded into the same codewords.

In the experimentation, we collect the Web pages in English and Chinese from Open Directory Project for the data sets. Five binary and three multi-class tasks have been set up. Our method gives significant improvement against several existing classifiers, and converges very well.

In summary, in this paper we make the following contributions:

- In addressing the cross-language Web page classification problem, we observe that there is a common part of Chinese and English documents and develop a novel method for addressing the problem of topic drifts in Chinese and English documents, thus improving the classification performance.
- We propose to handle noisy features and different features problem in cross-language Web-page classification by the information bottleneck technique. This method allows the common part in the two languages to be extracted and used for classification, despite their differences. We show experiments on real English and Chinese Web documents. Our method is shown to improve other classification algorithms.

The rest of our paper is organized as follows: In Section 2 we briefly discuss the related work. Following the basic concepts reviewed in Section 3, we introduce the *information bottleneck* theory in Section 4. Section 5 describes our proposed method in details. The experiments and results are presented in Section 6. In the end, we conclude this paper with future work discussion in Section 7. The detailed proofs to the lemmas and theorems in this paper will be given in Appendix.

2. RELATED WORK

In this section, we review several prior researches mostly related to our work, including traditional classification, cross-language classification and information theoretic learning.

2.1 Traditional Classification

The traditional classification formulation is built on the foundation of statistical learning theory. Two schemes are generally considered, where one is *supervised classification* and the other is *semi-supervised classification*. Supervised classification focuses on the case where the labeled data are sufficient, and where the learning objective is to estimate a function that maps examples to class labels using the labeled training instances. Examples of supervised classification algorithms include decision trees [25], K nearest neighbor methods [6], naive Bayes classifiers [17], support vector machines [5], and so on.

Semi-supervised classification [32] addresses the problem that the labeled data are too few to build a good classifier. It makes use of a large amount of unlabeled data, together with a small amount of the labeled data to enhance the classifiers. Many semi-supervised learning techniques have been proposed, e.g., co-training [4], EM-based methods [21], transductive learning [13] etc.

As we claimed, there are two main difficulties in the cross-language classification, namely errors by translation and bias by topic drifts, which traditional classifiers cannot handle well. Our proposed method tries to tackle these difficulties via the *information bottleneck* technique.

2.2 Cross-Language Text Classification

There are several research works addressing the cross-language classification problem. Bel et al. [3] studied English-Spanish cross-language classification problem. Two scenarios are considered in their work. One scenario assumes to have training documents in both languages. The other scenario is to learn a model from the text in one language and classify the data in another language by translation. In our work, we focus on the second scenario. [26] gave good empirical results on English-Italian cross-language text categorization using an EM-based learning method. Note that to avoid trivial partitions, it applies feature selection before each iteration. [23] employed a general probabilistic English-Czech dictionary to translate Czech text into English and then classified Czech documents using the classifier built on English training data. Other cross-language text classification research include [11] (English-Spanish), [18] (English-Japanese) etc, to be mentioned. In addition to text catego-

rization, there are some other specific cross-language applications: Named Entity Recognition [30], Question Answering [8], etc.

However, most existing algorithms are based on traditional supervised or semi-supervised classification techniques. As we stated, there are two difficulties in the cross-language classification, the translation error and topic drift, which lead to difference in distributions between the Web pages in two languages. Since traditional supervised classification techniques assume identical distribution for training and test data and in the cross-language setting this assumption is hardly met, most existing algorithms will not cope with the cross-language text classification well. In this work, our method tries to handle the Chinese-English cross-language categorization problem by information bottleneck. We will show that our algorithm can better alleviate the impact of translation error and topic drift, and improve the (English to Chinese) cross-language classification performance against existing methods.

2.3 Information Theoretic Learning

Another related research is information theory based learning. Information theory is widely used in machine learning, e.g. decision tree [25], feature selection [31], etc.

The *information bottleneck* theory (IB) was first proposed by Tishby et al. [29]. They constructed a model that uses information theory to solve the clustering problem. In their work, a rate distortion function is introduced as a loss function. They also presented a converging iterative algorithm for this self-consistent determination problem. After that, a lot of interesting works have been conducted, c.f. [27]. As known, IB is an information theoretic formulation for clustering problem while maximum likelihood of mixture models is a standard statistical method to clustering. Interestingly, Slonim and Weiss [28] have proved that under a certain mapping, these two approaches are strongly related. Moreover, when input data is large enough, they are statistically equivalent.

Several extensions to information bottleneck method have been investigated recently. [9] proposed a word clustering method which minimizes the loss in mutual information between words and class-labels, before and after clustering. Using similar strategy, mutual information based [10] and Bregman divergence based [2] co-clustering were proposed.

In contrast to these works, we focus on solving the cross-language classification problem via an information theoretic approach. More specifically, the IB technique is used to mine the common part of the pages in different languages for classification.

3. PRELIMINARY

In this section, some preliminary knowledge in the information theory is briefly introduced, including information entropy, mutual information and Kullback-Leibler divergence [14]. For more details, please refer to [7].

The term *entropy* is used to measure the uncertainty associated with a random variable X. Formally,

$$H(X) = -\int_x p(x)\log(p(x))\mathrm{d}x, \tag{1}$$

where x enumerates each value X may take. In the communication system, the entropy quantifies the information contained in a piece of data, i.e. its minimum average message length in bits. This means the best possible lossless data compression is limited to this measure.

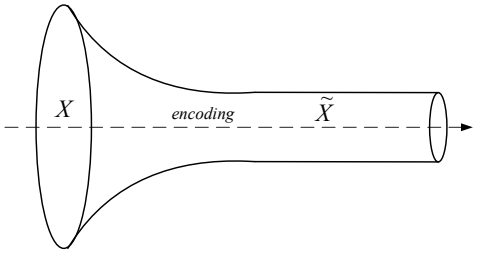

Figure 2: The information bottleneck. Here, X is the signals to be encoded, and $\tilde{X}$ is the codewords. In classification, X is the set of instances, and $\tilde{X}$ is the prediction labels.

Let X and Y be random variable sets with a joint distribution $p(X,Y)$ and marginal distributions $p(X)$ and $p(Y)$. The *mutual information* $I(X;Y)$ is defined as

$$I(X;Y) = \int_x\int_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y. \tag{2}$$

The mutual information is a measure of the dependency between random variables. It is always non-negative, and it is zero if and only if the variables are statistically independent. Higher mutual information values indicate more certainty that one random variable depends on another.

The mutual information is related to the *Kullback-Leibler (KL) divergence* or *relative entropy* measures, defined for two probability mass functions $p(x)$ and $q(x)$,

$$D(p||q) = \int_x p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x, \tag{3}$$

where $q(x)$ is a reference distribution. The KL-divergence can be considered as a kind of a distance between the two probability distributions, although it is not a real distance measure because it is not symmetric. In addition, KL-divergence is always non-negative due to the Gibbs' inequality [7].

4. INFORMATION BOTTLENECK

In this section, we present the basic concepts for information bottleneck, and show how it can be applied to cross-language classification.

4.1 Basic Concepts

The *information bottleneck* (IB) method is a distributional learning algorithm proposed by Tishby et al. [29]. In this theory, the clustering and classification problems can be treated as a coding process. Let X be the signals to be encoded, and $\tilde{X}$ be the set of codewords. In classification, the codewords $\tilde{X}$ is defined as class labels, and then classifying X can be seen as using $\tilde{X}$ to encode X. Usually $\tilde{X}$ is not able to contain as much information as X. Therefore, when the signals X is encoded to codewords $\tilde{X}$, part of the information contained by X will be lost. This can be imaged as the information in X passing through a bottleneck $\tilde{X}$, as shown in Figure 2, and thus $\tilde{X}$ is called *information bottleneck*.

A good clustering should guarantee the encoding effectiveness (the lower rate the better) as well as meaningful information, which can be formulated as a trade-off objective $I(\tilde{X};X)-\beta I(\tilde{X};Y)$ [29], where Y is the feature set with respect to X. Note that the coding rate $I(\tilde{X};X)$ in classification is no so important as in clustering problems, since classification focuses on prediction accuracy. Therefore, in this task, we need to mainly concentrate on improving the classification accuracy. So that, $-I(\tilde{X};Y)$ is optimized instead of $I(\tilde{X};X)-\beta I(\tilde{X};Y)$, as the coding rate $I(\tilde{X};X)$ is ignored. Moreover, in order to make the optimization easier, in this work, we optimize $I(X;Y)-I(\tilde{X};Y)$ instead of $-I(\tilde{X};Y)$. We will show the optimization details in the next section. Note that, $I(X;Y)$ is a constant, when X and Y are fixed, and thus optimizing $I(X;Y)-I(\tilde{X};Y)$ is equivalent to optimizing $-I(\tilde{X};Y)$.

The meaning of objective function $I(X;Y)-I(\tilde{X};Y)$ can be also understood in another way. In classification, the instances are described by the features, and thus there should be mutual information between data instances and their features, i.e. $I(X;Y)>0$. In addition, the category information is also described by the features, and thus $I(\tilde{X};Y)>0$. Therefore, the information in X and $\tilde{X}$ is contained in forms of $I(X;Y)$ and $I(\tilde{X};Y)$. Note that $I(X;Y)$ is always greater than or equal to $I(\tilde{X};Y)$, which will be derived in Lemma 1 in Section 5.2. $I(X;Y)-I(\tilde{X};Y)>0$ means there is some *loss in mutual information* after categorization. To sum up, a good categorization should keep the mutual information between data and features, and minimize the information loss. In other words, $I(\tilde{X};Y)$ should be close to $I(X;Y)$. Therefore, in the information bottleneck classification setting, the quality of the categorization should be judged by the *loss in mutual information* between the original instances and categorized instances. The following remark lays our the objective function formally.

REMARK 1. *In the information bottleneck (IB) classification setting, a qualified hypothesis $h: X \mapsto \tilde{X}$ approximately satisfies:*

$$h = \arg\min_{h^*} \Big(I(X;Y) - I(\tilde{X};Y) \Big) . \quad (4)$$

4.2 Applying to Cross-Language Classifier

Before the formal problem formulation, we would like to present a brief idea to address the cross-language classification problem by information bottleneck. Suppose we have the union set X of English and Chinese Web pages, Y is the set of words contained in X, and $\tilde{X}$ stands for the class labels.

It is observed that there is a common part of English and Chinese web pages, which share the similar semantic information. This observation inspired our work. If the texts are put through the "bottleneck", the labeled data (English Web pages) will, to some extent, guide the classification on Chinese pages by encoding the common part into the same codewords. It is because that during the information theoretic compression, IB tends to encode the similar pages into the same code (label) in that it can reduce the code length without loss of much information. If one Chinese page is in the common part, i.e. similar to a labeled English page, this page will be classified into the same category as that of the English one.

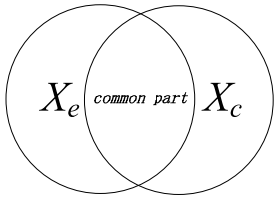

Figure 3: The cross-language classification problem. Here, X_e is the set of the Web pages in English and X_c is the set of the Web pages in Chinese. It is assumed that there is some common part between English and Chinese Web pages.

To summarize, it is noted that cross-language Web pages are carrying a common part of semantic information. According to the information bottleneck theory, this part of information is expected to help encode similar Web pages in different languages since it is highly relevant information to class labels and also is contained in the Web pages in both languages. Therefore, it is clear that the information bottleneck technique is suitable to address the cross-language classification problem.

5. CROSS-LANGUAGE CLASSIFIER VIA INFORMATION BOTTLENECK

In this section, the problem is carefully formulated and an objective function is proposed to build a classification model. The classification algorithm (IB) is then presented. Also, the convergence and time complexity are theoretically analyzed.

5.1 Problem Formulation

Let X_e be the set of the Web pages in English with class labels and X_c be the set of the Web pages in Chinese without labels. Usually, it is assumed that the English Web pages X_e and Chinese Web pages X_c share some common information with each other, as shown in Figure 3. The feature space for X_e is the English words Y_e and for X_c is the Chinese words Y_c. The objective is, we are trying to classify the pages in X_c into C, the predefined class label set, which is the same for the pages in X_e. To sum up, labeled training documents are available only in one language, and we want to estimate a hypothesis $h : X_c \mapsto C$ which classifies documents written in another language. This is called cross-language text classification.

In contrast to traditional text categorization, where the training and test pages are in the same language, cross-language text classification requires that the training and test data should be unified into one single feature space. Otherwise, it is not possible for existing machine learning techniques to get the results. The most common way is to translate the pages in one language into the other one. However, it is noticed that this common approach will bring the error and bias for further classification. Linguistically speaking, machine translation technique is far from satisfactory. What is worse, the simple translation preprocessing does not give the accurate information. The topics of original data may drift under translation. Empirically, these claims were justified. The experimental details are presented in Section 6.1.2.

5.2 Objective Function

As stated in Section 4.2, we propose to address the problems via the information bottleneck technique. The test Web page set is translated to English, denoted as X_c^T. Let $X = X_e \cup X_c^T$, as the original signal for the bottleneck. The class label set is the output of the bottleneck $\tilde{X}$. Y is referred to the features of all the documents. To fully utilize the common part for classification, we defined the objective function as

$$I(X;Y) - I(\tilde{X};Y), \tag{5}$$

which is exactly the objective function in Remark 1. Note that $I(X;Y) - I(\tilde{X};Y)$ is always non-negative, which will be derived by Lemma 1. Based on Remark 1, the objective function value should be minimized, since we aim to draw $I(\tilde{X};Y)$ close to $I(X;Y)$.

Before proposing the optimization approach for minimizing the objective function, we first define some notations used in subsequent analysis.

DEFINITION 1. *We use $\tilde{X} = \{\tilde{x}\}$ to denote a categorization of X for the hypothesis h, where $\tilde{x} = \{x' | h(x') = h(x)\}$. Clearly, $|\tilde{X}|$ is equal to $|C|$, since h maps the instances in X to the class-labels in C.*

DEFINITION 2. *The joint probability distribution of X and Y under the categorization $\tilde{X}$ is denoted by $\tilde{p}(X,Y)$, where*

$$\tilde{p}(x,y) = p(\tilde{x},y)p(x|\tilde{x}) = p(\tilde{x},y)\frac{p(x)}{p(\tilde{x})}, \tag{6}$$

where $x \in \tilde{x}$, and $p(x|\tilde{x}) = \frac{p(x)}{p(\tilde{x})}$ since x totally depends on $\tilde{x}$.

In the following, we will transform the objective function in Equation (5) into another representation by KL-divergence [14].

LEMMA 1. *For a fixed categorization $\tilde{X}$, we can write the objective function in Equation (5) as*

$$I(X;Y) - I(\tilde{X};Y) = D(p(X,Y)||\tilde{p}(X,Y)), \tag{7}$$

where $D(\cdot||\cdot)$ is the KL-divergence defined as Equation (3).

Note that, based on the non-negativity of KL-divergence, the objective function $I(X;Y) - I(\tilde{X};Y)$ should always be non-negative.

5.3 Optimization

From Equation (7), it is found that the loss in mutual information in the objective function equals to the KL-divergence between $p(X,Y)$ and $\tilde{p}(X,Y)$. To minimize the objective function in Equation (5), we need only to find a categorization $\tilde{X}$ which minimizes the KL-divergence value

$$D(p(X,Y)||\tilde{p}(X,Y)). \tag{8}$$

However, the objective function in Equation (7) is in the joint probability form that is difficult to be optimized. Now, we are to rewrite it into a conditional probability form, which will facilitate our algorithm to reduce the objective function value.

LEMMA 2. *The objective function in Equation (7) can be expressed by a conditional probability form as*

$$D(p(X,Y)||\tilde{p}(X,Y)) = \sum_{\tilde{x}\in\tilde{X}}\sum_{x\in\tilde{x}} p(x)D(p(Y|x)||\tilde{p}(Y|\tilde{x})). \tag{9}$$

Lemma 2 gives a straightforward explanation for cross-language text classification via the information bottleneck technique. If one Chinese page is more similar to the English pages of one class $\tilde{x}$, i.e. the distance $D(p(Y|x)||\tilde{p}(Y|\tilde{x}))$ is the smallest, assigning this page to $\tilde{x}$ will lead to a lower value of the objective function. Then it is desirable during the optimization process, which means the common part between English and Chinese Web pages work as stated in Section 4.2.

Also, Lemma 2 provides an alternative way to reduce the objective function value. From Equation (9), we know that minimizing $D(p(Y|x)||\tilde{p}(Y|\tilde{x}))$ for a single instance x could reduce the global objective function $D(p(X,Y)||\tilde{p}(X,Y))$. As a result, if we iteratively optimize the corresponding $D(p(Y|x)||\tilde{p}(Y|\tilde{x}))$ for each instance x, the objective function will decrease monotonically. Thus, based on Lemma 2, the information bottleneck cross-language text classification (IB) is derived as in Algorithm 1.

Algorithm 1 The Cross-Language Text Classification (IB) Algorithm

Input: English Web pages X_e; Chinese Web pages X_c; an existing translator T; the number of iterations N.
Output: the final hypothesis $h_f : X_c \cup X_e \mapsto C$.

1: Translate X_c into English: $X_c^T = T(X_c)$. Let $X = X_e \cup X_c^T$.
2: Train an initial hypothesis $h^{(0)}$ based on X_e by supervised learning method (e.g. naive Bayes classifiers [17]).
3: Initialize the probability distribution $\tilde{p}^{(0)}$ based on $h^{(0)}$ and Equation (6).
4: **for** $t = 1, \ldots, N$ **do**
5: **for** each $x \in X_c^T$ **do**
6: $h^{(t)}(x) = \arg\min_{\tilde{x}\in\tilde{X}} D(p(Y|x)||\tilde{p}^{(t-1)}(Y|\tilde{x}))$
7: **end for**
8: **for** each $x \in X_e$ **do**
9: $h^{(t)}(x) = h^{(t-1)}(x)$
10: **end for**
11: Update $\tilde{p}^{(t)}$ based on $h^{(t)}$ and Equation (6).
12: **end for**
13: Return $h^{(N)}$ as the final hypothesis h_f.

In Algorithm 1, in each iteration, the algorithm keeps the prediction labels for the English Web pages X_e unchanged since their *true* labels are already known, while choosing the best category $\tilde{X}$ for each data instance x in X_c^T to minimize the function $D(p(Y|x)||\tilde{p}^{(t-1)}(Y|\tilde{x}))$. As we have discussed above, this process is able to decrease the objective function in Equation (7). The whole algorithm is illustrated in Figure 4.

5.4 Convergence

Since our algorithm IB is iterative, it is necessary to discuss its property of convergence. The following theorem shows that the objective function in our algorithm monotonically decreases, which establishes that the algorithm converges eventually.

THEOREM 1. *The objective function in Equation (7) monotonically decreases in each iteration of Algorithm IB.*

$$D(p(X,Y)||\tilde{p}^{(t)}(X,Y)) \geq D(p(X,Y)||\tilde{p}^{(t+1)}(X,Y)). \tag{10}$$

Note that, although the algorithm is able to minimize the objective function value in Equation (5), it is only able to

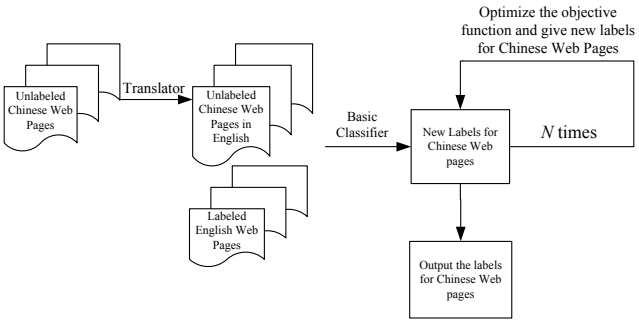

Figure 4: The scheme of the IB-based cross-language classification algorithm.

find a locally minimal one. Finding the global optimal solution is NP-hard. From Theorem 1, we can straightforwardly derived that the algorithm IB converges in a finite number of iterations, since the hypothesis space is finite.

5.5 Computational Complexity

Regarding the computational cost for IB, suppose the nonzeros in $p(X,Y)$ is N. In each iteration, IB needs to calculate $h^{(t)}$ in $O(|C| \cdot N)$ and update $\tilde{p}^{(t)}(Y|\tilde{X})$ in $O(|C| \cdot |Y|)$. Therefore, the time complexity of IB is $O(|C| \cdot (|Y| + N))$ as a result. Usually, $|C|$ is not large and could be considered as a constant, while $|Y|$ is usually not larger than N. Thus, the time complexity of IB is $O(N)$ in general. Thus, our algorithm IB has good scalability, and is capable for large data sets.

6. EXPERIMENTS

In this section, we evaluate our cross-language classification algorithm based on information bottleneck, and compare our algorithm with several state-of-art supervised and semi-supervised classifiers.

6.1 Data Sets

We conduct our evaluation on the Web pages crawled from the Open Directory Project (ODP) [22] during August 2006. Each Web page in ODP was classified by human experts into 17 top level categories (`Arts`, `Business`, `Computers`, `Games`, `Health`, `Home`, `Kids and Teens`, `News`, `Recreation`, `Reference`, `Science`, `Shopping`, `Society`, `Sports`, `Regional`, `Adult` and `World`). We removed the `Regional` category because the Web pages in the `Regional` category are also in other categories. The Web pages in the `Adult` have not been crawled by our crawler, because most of them are banned by our internet service provider, and thus the `Adult` category is not included in our data collection. Moreover, the Web pages in the `World` category are in the languages other than English. We selected all the Chinese pages from the `World` category as Chinese test data. For Chinese Web pages, there are also 14 top categories each of which can be mapped to a top category in the English ODP. Therefore, we have 14 categories for both Chinese and English Web pages in this experiments. Figure 1 represents the detailed description for our data collection. From the table, it can be seen that, the number of English labeled Web pages is much larger than that of Chinese ones in ODP, which indicates English labeled Web pages (1,271,106) is much more abundant than Chinese ones (18,243).

Category	Chinese Web Pages	English Web Pages
Arts	1,942	186,307
Business	6,503	203,569
Computers	1,907	102,571
Games	296	39,269
Health	518	47,607
Home	203	23,117
Kids and Teens	292	27,323
News	359	96,510
Recreation	681	77,901
Reference	2,338	48,231
Science	914	75,434
Shopping	488	86,736
Society	1,481	185,466
Sports	321	71,065
Total	18,243	1,271,106

Table 1: The descriptions for all the categories in ODP, including Chinese and English ones. Note that, all the Chinese Web pages as well as their category labels were translated into English using Google Translator.

Data Sets	Categories
Games vs News	Games, News
Arts vs Computers	Arts, Computers
Recreation vs Science	Recreation, Science
Computers vs Sports	Computers, Sports
Reference vs Shopping	Reference, Shopping
3 Categories	Arts, Computer, Society
4 Categories	Business, Health, News, Sports
5 Categories	Games, Home, News, Shopping, Sports

Table 2: The composition for each data sets.

6.1.1 Data Preparation

Data preprocessing has been applied to the raw data. First, all the Chinese Web pages were translated by Google Translator [1]. Then, we converted all the letters to lower cases, and stemmed the words using the Porter's stemmer [24]. After that, stop words were removed. In order to reduce the size of the feature space, we used a simple feature selection method, *document frequency* (DF) thresholding [31], to cut down the number of features, and speed up the classification. Based on [31], DF thresholding, which has comparable performance with *information gain* (IG) or CHI, is suggested since it is simplest with lowest cost in computation. In our experiments, we set the DF threshold to 3. After feature selection, the vocabulary size becomes 512,896.

In order to evaluate our cross-language classifier, we set up eight cross-language classification tasks. Five of them are binary classification tasks, and others are for multiple-class classification. Table 2 presents the detailed composition for each classification task.

6.1.2 Cross-Language Topic Drift

We extracted the most frequent features in the Chinese and English Web pages for each of the 14 ODP categories, and found the frequent features in the Chinese and English Web pages are quite different, although they share some

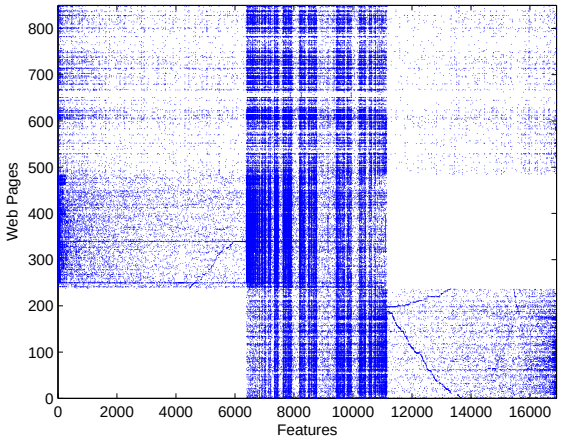

Figure 5: The instance-feature co-occurrence density for the `Games vs News` data set.

part. Table 3 presents the top 5 frequent features in the Chinese and English Web pages for each of the 14 ODP categories. We believe there is topic drift between Chinese and English Web pages. For example, there are several Chinese words which are frequently appears in the Chinese Web pages, such as "qiyuan", "mufurong", "pingqiu" and so on. In the `Games`, "qiyuan", one of the famous online game in China, is the most frequent keyword, while it hardly appears in the English Web pages. This is due to the difference in culture between the Chinese and western societies. "pingqiu", which means "draw" in English, hardly appears in English Web pages, because the translator fails to provide a mapping between "draw" and it. The observations demonstrate the claim we made previously that there are two main obstacles for the cross-language classification: one is the difference in topic focus between the two languages; the other is the translation error.

6.1.3 Density Analysis

Figure 5 shows the instance-feature co-occurrence distribution on the `Games vs News` data set. In this figure, documents 1 to 484 are from X_e, while documents 485 to 853 are from X_c. Within a data set, X_e or X_c, the documents are ordered by their categories (`Games` or `News`). The words are sorted by $n_g(w)/n_n(w)$, where $n_g(w)$ and $n_n(w)$ represent the number of word positions w appears in `Games` and `News` documents, respectively. From Figure 5, it can be found that the distributions of English and Chinese Web pages are somewhat different, however the figure also shows large commonness exists between the two data sets. The density divergence between two data sets in the figure makes the cross-language classification difficult, because most classification techniques rely on the basic assumption that the training data should be drawn from the same distribution as the test data. However, the common part between the two data sets can help increase the feasibility of the classification.

6.2 Comparison Methods

In this experiments, we compare our information bottleneck cross-language classifier (IB) with several state-of-art classification algorithms to show the advantages of our algorithm.

We take the supervised classification algorithms to be the baseline methods. Naive Bayes classifiers (NBC) [17] and support vector machines (SVM) [5] are evaluated in the experiments. They are trained on X_e and tested on X_c^T. Transductive support vector machines (TSVM) [13] is also introduced as comparison semi-supervised learning methods, which take both labeled X_e and unlabeled X_c^T for training and X_c^T for testing.

For implementation details, TF-IDF is used for feature weighting when training support vector machines (SVM) [5] and transductive support vector machines (TSVM) [13]. TF is used for feature weighting when training naive Bayes classifier (NBC) [17] and our information bottleneck based cross-language classification algorithm (IB).

SVM and TSVM are implemented by SVM^{light} [12] with default parameters (linear kernel). For more details about SVM and TSVM, please refer to [5] and [13]. NBC and IB are implemented by ourselves. The initial categorizations for IB are given by NBC.

6.3 Classification Performance

We now present the classification performance for each comparison methods, and show advantages of our information bottleneck cross-language classifier IB.

6.3.1 Evaluation Metrics

The metrics used in this experiments are macro-average precision, recall and F_1-measure. Let f be the function which maps from document d to its true class label $c = f(d)$, and h be the function which maps from document d to its prediction label $c = h(d)$ given by the classifiers. The macro-average precision P and recall R are defined as

$$P = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{|\{d|d \in X_c \wedge h(d) = f(d) = c\}|}{|\{d|d \in X_c \wedge h(d) = c\}|}, \quad (11)$$

$$R = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{|\{d|d \in X_c \wedge h(d) = f(d) = c\}|}{|\{d|d \in X_c \wedge f(d) = c\}|}. \quad (12)$$

F_1-measure is a harmonic mean of precision and recall defined as follows

$$F_1 = \frac{2PR}{P+R}. \quad (13)$$

6.3.2 Experimental Results

Table 4 presents the performance on each binary classification data set given by NBC, SVM, TSVM and our algorithm IB. The implementation details of the algorithms have already been presented in the last subsection. The evaluation metrics are macro-average precision, recall and F_1-measure, of which we have just given the definitions. From the table, we can see that IB significantly improves the other three methods. Although SVM and TSVM is slightly better than IB on the `Arts vs Computers` data set, IB is still comparable. But, on some of the other data sets, e.g. `Computers vs Sports` and `Reference vs Shopping`, both SVM and TSVM fail, while IB is much better than the two discriminative methods. In addition, NBC is always worse than IB, but never fails a lot. In average, IB gives the best performance in all the three evaluation metrics.

Table 5 presents the performance on each multiple-class classification data set given by NBC and our algorithm IB.

Category	Chinese Web Pages	English Web Pages
Arts	giotto, tugen, penchant, ashima, banzai	paean, base, dvdlaser, crew, taglin
Business	congeni, nanci, decre, wallk, darshan	natstat, kazaa, bcm, aanspreken, wct
Computers	volp, uptim, screenshot, malcolm, datastorag	volp, easyb, letterhack, grundriss, wsmcafe
Games	qiyuan, vierni, vernier, firstyeargirl, kangderong	accident, enix, impress, rifl, freeenergynew
Health	neurosyphili, maximowicziana, carboyhdr, podophyllin, interpol	finespun, stadium, linear, shyamalan, ryder
Home	banlan, xero, bcsahin, mufurong, prestonwood	machist, evita, beradino, bakk, fudoh
Kids and Teens	head, yangpu, ashaar, geetanjali, urdupoetri	pact, isch, argo, quem, melanesia
News	uppercut, muham, readjust, pulverul, dovic	narr, hume, mujer, dude, gif
Recreation	frauenhof, fatehpur, xingyuncao, nakedpoetri, orchha	behoof, heepster, bonafid, tiltrecord, kapil
Reference	filmcent, pessim, cold, seed, farm	platitudin, waggon, blankli, dfee, quotearch
Science	applaud, zhouyulin, flask, middlepillar, modarr	frobozzica, exploratori, forbrydelsen, kashyyk, pallot
Shopping	lcjzl, lashel, aubrac, roozi, ebullit	scherick, tricia, strewn, caryn, glenda
Society	cass, indispens, buddhist, tahm, trod	wallpap, hornadai, lafort, obstat, duranczyk
Sports	parama, cow, pingqiu, nesi, shiduotangbei	shinobi, jumbl, shirt, invert, sould

Table 3: The most frequent (stemmed) features in the Chinese and English Web pages for each ODP category. All the Chinese Web pages have already been translated into English using Google Translator.

Data Set	Precision				Recall				F_1-Measure			
	NBC	SVM	TSVM	IB	NBC	SVM	TSVM	IB	NBC	SVM	TSVM	IB
Games vs News	0.749	0.737	0.747	0.798	0.747	0.739	0.779	0.817	0.748	0.738	0.762	0.807
Arts vs Computers	0.731	0.783	0.768	0.767	0.728	0.801	0.785	0.782	0.730	0.792	0.776	0.774
Recreation vs Science	0.836	0.874	0.883	0.903	0.832	0.877	0.891	0.906	0.834	0.876	0.887	0.905
Computers vs Sports	0.783	0.669	0.611	0.844	0.800	0.840	0.759	0.873	0.792	0.745	0.677	0.858
Reference vs Shopping	0.911	0.743	0.650	0.929	0.827	0.858	0.766	0.859	0.867	0.797	0.703	0.893
Average	0.802	0.761	0.732	**0.848**	0.787	0.823	0.796	**0.847**	0.794	0.790	0.761	**0.847**

Table 4: Macro-average precision, recall and F_1-measure for each classifier on each binary classification data set. The training documents for NBC, SVM and IB are X_e and the test data for them are X_c^T. TSVM is trained on labeled X_e and unlabeled X_c^T, and tested on X_c^T.

Data Set	Precision		Recall		F_1-Measure	
	NBC	IB	NBC	IB	NBC	IB
3 Categories	0.647	0.661	0.630	0.665	0.638	0.663
4 Categories	0.445	0.592	0.570	0.648	0.500	0.619
5 Categories	0.550	0.608	0.488	0.582	0.517	0.627
Average	0.547	**0.620**	0.563	**0.632**	0.552	**0.636**

Table 5: Macro-average precision, recall and F_1-measure for each classifier on each multiple-class classification data set. The training documents for NBC and IB are X_e and the test data for them are X_c^T. Note that, SVM and TSVM are not included here, because they are designed for binary classification.

SVM and TSVM are not included here because they are designed for binary classification, and cannot cope well with the multiple-class classification problem. In this table, IB still gives significant improvements against the baseline method NBC. Therefore, we believe our algorithm IB is not only effective for English-Chinese cross-language classification, but also extensible for multiple-class classification.

6.4 Convergence

Since our algorithm IB is an iterative algorithm, an important issue for IB is the convergence property. Theorem 1 has already proven the convergence of IB theoretically. Now, let us empirically show the convergence property of IB. Figure 6 shows the test error rate curves as functions for each iteration on three data sets, `Arts vs Computers`, `Recreation vs Science` and `3 Categories`. From the figure, it can be seen

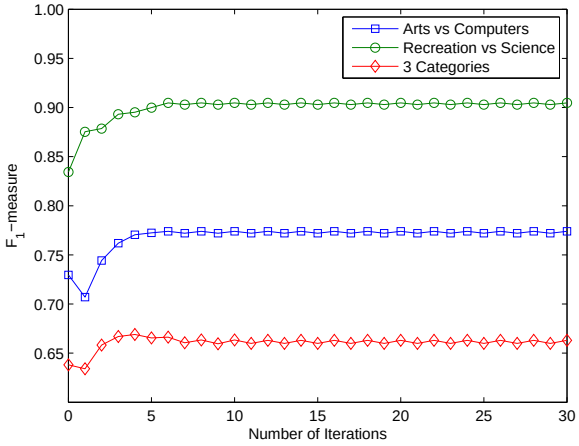

Figure 6: The F_1-measure curves after each iterations on three data sets `Arts vs Computers`, `Recreation vs Science` and `3 Categories` respectively.

that IB always achieves almost convergence points within 10 iterations. This indicates that IB converges very fast. We believe that 10 iterations is empirically enough for IB.

7. CONCLUSIONS AND FUTURE WORK

The tremendous growth of the World Wide Web in China has raised the need for classifying and organizing Chinese Web space via classification techniques. In this paper we

put forward a technique for the Chinese Web mining task to exploit the abundant labelled information in English. In particular, we have developed a novel method known as the information bottleneck technique to address the topic drift and different feature-space problems across two languages. Our method brings out a common part between the Chinese and English Web pages, which can be used to encode similar pages in different languages into the same codewords (class labels). An iterative algorithm is presented to optimize the objective function and therefore solve this problem. The experimental results show that our method can effectively improve existing methods in general, including five binary and three multi-class problems.

To extend our work, we wish to modify our method to achieve a global optimal value. It is also interesting to conduct more experiments in other language pair (e.g. French vs English, which does not suffer the word segmentation problem). Moreover, our method has the potential to be effective for the cross-language information retrieval problem.

8. ACKNOWLEDGEMENTS

Qiang Yang would like to thank the support of Hong Kong RGC Grant 621307. We also thank the anonymous reviewers for their great helpful comments.

9. REFERENCES

[1] Google translator. `http://www.google.com/language_tools`.

[2] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 509–514, 2004.

[3] N. Bel, C. Koster, and M. Villegas. Cross-Lingual Text Categorization. *Proceedings ECDL*, 200:126–139, 2003.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[5] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[8] M. Day, C. Ong, and W. Hsu. Question Classification in English-Chinese Cross-Language Question Answering: An Integrated Genetic Algorithm and Machine Learning Approach. In *Proceedings of the 2007 IEEE International Conference on Information Reuse and Integration*, pages 203–208, 2007.

[9] I. S. Dhillon, S. Mallela, and R. Kumar. Enhanced word clustering for hierarchical text classification. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 191–200, 2002.

[10] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[11] A. Gliozzo and C. Strapparava. Cross language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora. *Proc. of the ACL Workshop on Building and Using Parallel Texts (in conjunction of ACL-05)*, 2005.

[12] T. Joachims. SVM light. *Software available at* `http://svmlight.joachims.org`.

[13] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, 1999.

[14] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[15] K. Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[16] D. D. Lewis. Reuters-21578 test collection. `http://www.daviddlewis.com/`.

[17] D. D. Lewis. *Representation and learning in information retrieval*. PhD thesis, Amherst, MA, USA, 1992.

[18] Y. Li and J. Shawe-Taylor. Advanced learning algorithms for cross-language patent retrieval and classification. *Information Processing and Management*, 43(5):1183–1199, 2007.

[19] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru. Automatic search engine performance evaluation with click-through data analysis. *Proceedings of the 16th international conference on World Wide Web*, pages 1133–1134, 2007.

[20] X. Ni, G. Xue, X. Ling, Y. Yu, and Q. Yang. Exploring in the weblog space by detecting informative and affective articles. *Proceedings of the 16th international conference on World Wide Web*, pages 281–290, 2007.

[21] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2-3):103–134, 2000.

[22] ODP. Open directory project. `http://www.dmoz.com/`.

[23] J. Olsson, D. Oard, and J. Hajič. Cross-language text classification. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 645–646, 2005.

[24] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[25] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[26] L. Rigutini, M. Maggini, and B. Liu. An em based training algorithm for cross-language text categorization. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 529–535, 2005.

[27] N. Slonim. The Information Bottleneck: Theory and Applications. *Unpublished doctoral dissertation, Hebrew University, Jerusalem, Israel*, 2002.

[28] N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. *Advances in Neural Information Processing Systems*, 15, 2002.

[29] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the Thirty-seventh Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[30] Y.-C. Wu, K.-C. Tsai, and J.-C. Yang. Two-pass named entity classification for cross language question answering. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 168–174, 2007.

[31] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of Fourteenth International Conference on Machine Learning*, pages 144–152, 1997.

[32] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, University of Wisconsin–Madison, 2006.

APPENDIX

In this appendix, we provide the detailed proof to Lemmas 1 and 2, and Theorem 1.

A. PROOF TO LEMMA 1

PROOF.

$$\begin{aligned}
&I(X;Y) - I(\tilde{X};Y)\\
=&\sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\sum_{x\in\tilde{x}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\\
&\quad-\sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\left(\sum_{x\in\tilde{x}}p(x,y)\right)\log\frac{p(\tilde{x},y)}{p(\tilde{x})p(y)}\\
=&\sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\sum_{x\in\tilde{x}} p(x,y)\log\frac{p(x,y)}{p(\tilde{x},y)\frac{p(x)}{p(\tilde{x})}}\\
=&\sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\sum_{x\in\tilde{x}} p(x,y)\log\frac{p(x,y)}{\tilde{p}(x,y)}\\
=&D(p(X,Y)||\tilde{p}(X,Y))\,.
\end{aligned}$$

□

B. PROOF TO LEMMA 2

PROOF.

$$D(p(X,Y)||\tilde{p}(X,Y)) = \sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\sum_{x\in\tilde{x}} p(x,y)\log\frac{p(x,y)}{\tilde{p}(x,y)}\,.$$

Since

$$\begin{aligned}
\tilde{p}(x,y) &= p(\tilde{x},y)p(x|\tilde{x}) = p(\tilde{x},y)\frac{p(x)}{p(\tilde{x})}\\
&= p(x)p(y|\tilde{x}) = p(x)\tilde{p}(y|\tilde{x})\,,
\end{aligned}$$

we have

$$\begin{aligned}
D(p(X,Y)||\tilde{p}(X,Y)) &= \sum_{\tilde{x}\in\tilde{X}}\sum_{y\in Y}\sum_{x\in\tilde{x}} p(x)p(y|x)\log\frac{p(x)p(y|x)}{p(x)\tilde{p}(y|\tilde{x})}\\
&= \sum_{\tilde{x}\in\tilde{X}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\tilde{p}(y|\tilde{x})}\\
&= \sum_{\tilde{x}\in\tilde{X}}\sum_{x\in\tilde{x}} p(x)D(p(Y|x)||\tilde{p}(Y|\tilde{x}))\,.
\end{aligned}$$

□

C. PROOF TO THEOREM 1

PROOF. Based on Lemma 2, we have

$$D(p(X,Y)||\tilde{p}^{(t)}(X,Y)) = \sum_{\tilde{x}:h^{(t)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\tilde{p}^{(t)}(y|\tilde{x})}\,.$$

From the Steps 6 and 9 in Algorithm 1,

$$h^{(t)}(x) = \begin{cases} \arg\min_{\tilde{x}\in\tilde{X}} D(p(Y|x)||\tilde{p}^{(t-1)}(Y|\tilde{x})) & x\in X_o\\ h^{(t)}(x) = h^{(t-1)}(x) & x\in X_i\end{cases}\,.$$

Thus,

$$\begin{aligned}
&D(p(X,Y)||\tilde{p}^{(t)}(X,Y))\\
\geq&\sum_{\tilde{x}:h^{(t)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\tilde{p}^{(t)}(y|h^{(t+1)}(x))}\\
=&\sum_{\tilde{x}:h^{(t+1)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\tilde{p}^{(t)}(y|\tilde{x})}\\
=&\sum_{\tilde{x}:h^{(t+1)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\left(\log p(y|x) + \log\frac{1}{\tilde{p}^{(t)}(y|\tilde{x})}\right)\,.
\end{aligned}$$

Here,

$$\begin{aligned}
&\sum_{\tilde{x}:h^{(t+1)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{1}{\tilde{p}^{(t)}(y|\tilde{x})}\\
=&\sum_{\tilde{x}:h^{(t+1)}}\left(\sum_{x\in\tilde{x}}\sum_{y\in Y} p(x)p(y|x)\right)\log\frac{1}{\tilde{p}^{(t)}(y|\tilde{x})}\\
=&\sum_{\tilde{x}:h^{(t+1)}}\tilde{p}^{(t+1)}(\tilde{x})\sum_{y\in Y}\tilde{p}^{(t+1)}(y|\tilde{x})\log\frac{1}{\tilde{p}^{(t)}(y|\tilde{x})}\\
\geq&\sum_{\tilde{x}:h^{(t+1)}}\tilde{p}^{(t+1)}(\tilde{x})\sum_{y\in Y}\tilde{p}^{(t+1)}(y|\tilde{x})\log\frac{1}{\tilde{p}^{(t+1)}(y|\tilde{x})}\,.
\end{aligned}$$

Note that, the last inequality follows by the non-negativity of the Kullback-Leibler divergence, that

$$\begin{aligned}
&\sum_{y\in Y}\tilde{p}^{(t+1)}(y|\tilde{x})\log\frac{1}{\tilde{p}^{(t)}(y|\tilde{x})} - \sum_{y\in Y}\tilde{p}^{(t+1)}(y|\tilde{x})\log\frac{1}{\tilde{p}^{(t+1)}(y|\tilde{x})}\\
&\qquad= D(\tilde{p}^{(t+1)}(Y|\tilde{x})||\tilde{p}^{(t)}(Y|\tilde{x}))\geq 0\,.
\end{aligned}$$

Thus,

$$\begin{aligned}
&D(p(X,Y)||\tilde{p}^{(t)}(X,Y))\\
\geq&\sum_{\tilde{x}:h^{(t+1)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\left(\log p(y|x) + \log\frac{1}{\tilde{p}^{(t+1)}(y|\tilde{x})}\right)\\
=&\sum_{\tilde{x}:h^{(t+1)}}\sum_{x\in\tilde{x}} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\tilde{p}^{(t+1)}(y|\tilde{x})}\\
=&D(p(X,Y)||\tilde{p}^{(t+1)}(X,Y))\,.
\end{aligned}$$

□